ARL-RP-0561 ● DEC 2015

US Army Research Laboratory

Studies in the Human Use of Controlled English

by David Mott and Erin Zaroukian

Reprinted from the International Technology Alliance Collaboration System [accessed 2015 Dec 2]. https://www.usukitacs.com/node/3067. Paper presented at the 2015 US/UK International Technology Alliance Fall Meeting; 2015 Sep 15–17; College Park, MD.

NOTICES

Disclaimers

ARL-RP-0561 ● DEC 2015

US Army Research Laboratory

Studies in the Human Use of Controlled English

by David Mott
Emerging Technology Services, IBM United Kingdom, Ltd
Hursley Park, Winchester, UK

Erin Zaroukian
Human Research and Engineering Directorate, ARL

Reprinted from the International Technology Alliance Collaboration System [accessed 2015 Dec 2]. https://www.usukitacs.com/node/3067. Paper presented at the 2015 US/UK International Technology Alliance Fall Meeting; 2015 Sep 15–17; College Park, MD.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED (From - To)**
December 2015	Reprint	October 2014–September 2015

4. TITLE AND SUBTITLE Studies in the Human Use of Controlled English	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) David Mott and Erin Zaroukian	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory ATTN: RDRL-HRS-E Aberdeen Proving Ground, MD 21005-5425	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-RP-0561
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Reprinted from the International Technology Alliance Collaboration System [accessed 2015 Dec 2]. https://www.usukitacs.com/node/3067. Paper presented at the 2015 US/UK International Technology Alliance Fall Meeting; 2015 Sep 15–17; College Park, MD.

14. ABSTRACT

Controlled English (CE) is intended to aid human problem solving processes when analysing data and generating high-value conclusions in collaboration with computer systems. It is therefore important to evaluate the use of CE by human users when performing such problem solving. This paper describes two different approaches to such an evaluation. The first approach is anonymous online experimentation, where the participant sees the simultaneous presentation of a visual diagram of a particular state of affairs (or "ground truth") and a CE rule, and determines whether the rule corresponds to the state of affairs. The second approach is to guide a user face-to-face to formulate free English sentences into CE to solve a logic problem. The paper describes both approaches and provides an informal analysis of the results to date.

15. SUBJECT TERMS

controlled natural language, human machine interaction, human-computer collaboration, problem solving, negation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Erin Zaroukian
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER (Include area code)**
Unclassified	Unclassified	Unclassified	UU	14	410-278-7556

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Studies in the Human Use of Controlled English

David Mott
Emerging Technology Services,
IBM United Kingdom Ltd
Hursley Park, Winchester, UK

Erin Zaroukian
Human Research & Engineering Directorate
US Army Research Laboratory
Aberdeen Proving Ground, MD, USA

***Abstract*— Controlled English (CE) is intended to aid human problem solving processes when analysing data and generating high-value conclusions in collaboration with computer systems. It is therefore important to evaluate the use of CE by human users when performing such problem solving. This paper describes two different approaches to such an evaluation. The first approach is anonymous online experimentation, where the participant sees the simultaneous presentation of a visual diagram of a particular state of affairs (or "ground truth") and a CE rule, and determines whether the rule corresponds to the state of affairs. The second approach is to guide a user face-to-face to formulate free English sentences into CE to solve a logic problem. The paper describes both approaches and provides an informal analysis of the results to date.**

I. Introduction

This paper reports work under the International Technology Alliance (ITA) on supporting collaborations of human and machine in the execution of problem-solving tasks, such as those faced by analysts when inferring high-value information from a variety of sources. These are complex cognitive tasks that require assumption making and reasoning based upon a "conceptual model" of the domain in which the analysis is taking place, and our research goal is to support users in such tasks by providing a language in which they may express their knowledge, concepts, rules and problem-solving strategies, called ITA Controlled English (CE) [1].

CE is a Controlled Natural Language, a subset of English, that is both human-readable and machine parseable, suitable for the expression of domain knowledge, concepts and reasoning, but also has a formal interpretation that is sufficiently unambiguous that a computer can interpret the input of the domain analysts and use it to perform inferencing. Central to the use of CE is a conceptual domain model, a structure that holds the users' knowledge (i.e. concepts, relationships, logical inferences, constraints, assumptions) of the domain in which the problem solving is to be undertaken. The analyst's problem solving strategies may also be represented in CE as ways of reasoning and of making assumptions, which vary based on the level of expertise of the analyst and the domain of analysis. The reasoning can be tracked through the rationale, showing how conclusions are dependent upon givens and assumptions.

We are researching the use of CE for supporting complex problem solving, as typified by the Analysis Game [2], a problem designed to teach analysts how to avoid analytic pitfalls. Whilst some success has been achieved in using CE reasoning to solve these tasks, it is important to determine whether CE can be taken up by more human users and applied to developing solutions for more complex problems. In order for CE to aid human problem solving, CE must be comprehensible to a human user, and it is important to understand the types of issues that users face when interacting with CE-based systems. Although some work has been done to objectively assess how well human users understand Controlled Natural Languages [3], this has not been done for ITA Controlled English. This paper reports some of the initial findings of two studies to explore the use of CE by non-specialist users. It is anticipated that the experience of exposing the CE reasoning technology to the users will provide useful feedback that can guide further research in the use of CE and in the types of human interfaces that could help users.

II. Two Approaches

The studies were developed during a collaboration between David Mott (IBM UK) and Erin Zaroukian (ARL) at IBM Hursley in early 2015 [4]. Two types of study were designed, both with the general purpose of discovering whether human users are able to use CE to perform reasoning, but each having a different philosophy and approach. In general terms these are:

- **<u>MTurk-style</u>**, an experiment where a user is invited to participate via Amazon's Mechanical Turk (MTurk), and where there is no direct interaction with an experimenter. The experiment is based upon the simultaneous presentation of a visual diagram of a "state of affairs" and a CE rule, and the user is asked whether the rule is consistent with the diagram.
- **<u>Face-to-Face</u>**, an evaluation where a user is guided in a face-to-face meeting with the evaluator to formulate one or more full English sentences into CE in order to solve a logic problem. As part of the study, the user is asked to provide comments on their problem solving, allowing access, in an informal way, to their cognitive reasoning.

The MTurk-style experiment was aimed at providing solid scientific data on the behaviour of a large range of users, whereas the face-to-face evaluation was aimed at providing informal information about the finer details of the user's cognitive and reasoning processes.

III. The MTurk Experiment

This section describes the development and implementation of a behavioural research paradigm to objectively assess how well human users understand Controlled Natural Languages. We demonstrate this paradigm with a case study: assessing different ways of asserting in CE that two entities are unique (e.g. "the thing A is not the thing B", "the thing A cannot be the thing B"). Using a methodology similar to Kuhn (2009) [3], study participants are presented with a CE rule and a diagram indicating the ground truth of relationships among the depicted

entities, and they must demonstrate comprehension (or lack thereof) by deciding whether the diagram is consistent with the CE rule. Preliminary results gathered through MTurk indicate that this type of task is feasible, both for an experimenter to implement and for a participant to complete.

The aim of this design is to assess whether particular types of CE statements are easier/harder to comprehend, allowing us to then make evidence-based recommendations to support changes or guidelines for CE and other Controlled Natural Languages.

A. *Design principles*

Some of the design principles for the experiment are: to be able to collect data quickly, with no experimenter intervention, and in a controlled and rigorous way; to avoid questions that have ambiguous answers; to avoid known difficulties in human understanding of logical statements and syllogistic reasoning; to focus on a task where there is linguistic choice in the way that facts are stated in free English and in CE, which has the potential to affect human construction of meaning from sentences.

B. *Method*

Participants – 45 participants were recruited through Amazon Mechanical Turk and were paid $.75 for their participation. Participants optionally provided their age, gender, background in logic/programming, and whether they were a native speaker of English.

Materials and procedures – participants were presented with a rule written in CE and a diagram, and they were asked to decide (Yes/No) if the diagram was consistent with the rule.

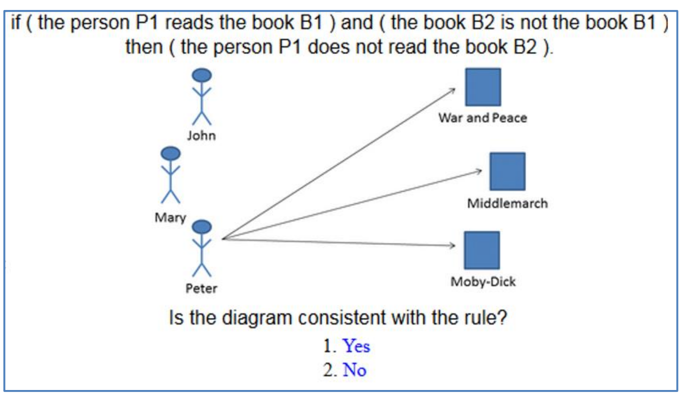

Fig. 1. Sample presentation to user via MTurk

Diagrams – The diagrams used in this study are modeled after 'ontographs' [5] and represent a closed world, in that if something is missing from the diagram it can be taken to be false. All diagrams in this study contain three people (John, Mary, Peter), three books (War and Peace, Middlemarch, and Moby-Dick), and reading relations represented by an arrow pointing from the reader to the book read. Four diagrams with relatively simple relations were used in practice, and four diagrams with relatively more complex relations were used in test.

Facts – Uniqueness, the contrast of interest in this study, was expressed in three ways, exemplified below:

"the person John is not the person Jim"
"the person John cannot be the person Jim"
"there is a person named John that is different to the person Jim"

Our intuitive hypothesis was that these expressions might cause difficulties in understanding to differing degrees.

Rules - All rules were of the form "if (STATEMENT) and (STATEMENT) then (STATEMENT)", where each STATEMENT was like the facts above, but had variables (such as P1, P2) instead of specific names (see Examples 1-4 below). A variable represents a specific but unknown object, and therefore a rule expresses a general pattern that can match many different situations. When uniqueness was expressed, it appeared as part of the antecedent "(STATEMENT) and (STATEMENT)".

To create variety in the items and discourage participants from developing superficial strategies, rules varied in:

- Whether uniqueness was expressed at all (some items had no uniqueness expression and were included as fillers, e.g. "if (the person P1 reads the book B1) and (the person P2 reads the book B2) then (the book B2 is the book B1).")
- Whether unbound variables were included (some items had unbound variables, i.e. variables in the conclusion that did not occur on the antecedent, and were included as fillers, e.g. "if (the person P1 reads the book B1) and (the person P2 cannot be the person P1) then (the person P2 reads the book B2)." where B2 is unbound).
- Whether people or books were expressed as unique (1–2 vs. 3–4 below)
- Whether the conclusion was positive or negative (1,3 vs. 2,4)
- Whether the order of antecedent conjuncts was regular (e.g. 1) or reversed e.g. "if (the book B2 is not the book B1) and (the person P1 reads the book B1) then (the person P1 does not read the book B2).")

Examples of the four main rule types are given below in CE, with "is not" as the uniqueness expression in the second antecedent statement. Each is followed by a natural-language paraphrase, though these were not available to participants.

1. if (the person P1 reads the book B1) and (the book B2 is not the book B1) then (the person P1 does not read the book B2). *'If a person reads a book, that person does not read any other book.'*
2. if (the person P1 reads the book B1) and (the book B2 is not the book B1) then (the person P1 reads the book B2). *'If a person reads a book, that person reads every other book too.'*
3. if (the person P1 reads the book B1) and (the person P2 is not the person P1) then (the person P2 does not read the book B1). *'If a person reads a book, no other person reads that book.'*
4. if (the person P1 reads the book B1) and (the person P2 is not the person P1) then (the person P2 reads the book B1). *'If a person reads a book, every other person reads that book too.'*

Procedure - Participants began with seven practice items, which introduced them to this statement–diagram paradigm and taught them how to read CE rules and interpret CE variables, but

which did not contain any uniqueness expressions. Participants were guided through how to solve four of the practice items, and for all practice items, after submitting a response, participants were told whether their response was correct or incorrect and were given an explanation of how to solve that item.

After the practice, participants saw 24 test items, of which 16 contained the contrast of interest and eight were fillers (described above). Each participant saw each of the four main rule types with each of the four diagrams: two which made it true, two which made it false. A Latin square design determined which uniqueness expression was used in each rule, and the number of regular/reversed antecedents was balanced within subjects.

C. Results

With fillers removed, mean accuracy was 0.747 (Standard Error = 0.031) with mean reaction time 12.802s (SE = 1.309s). The plots below show mean accuracy and response time per worker, with overall means as a gray–dashed line.

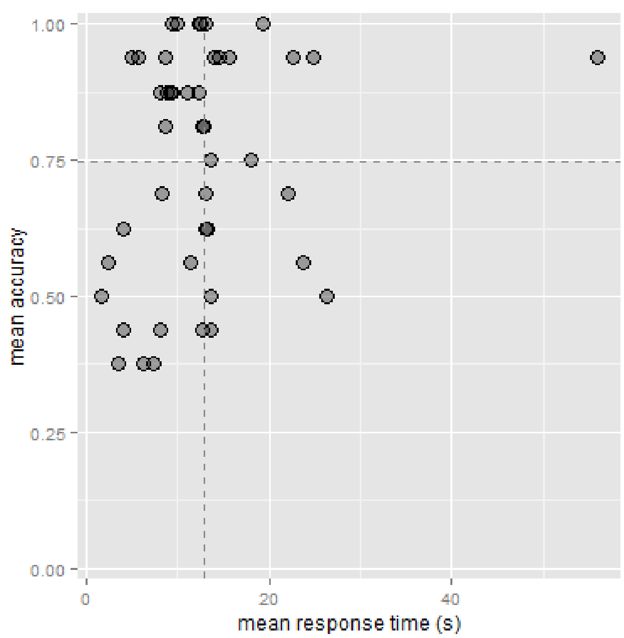

Fig. 2. Mean accuracy and response time for each worker

While there is little data from participants with "A Lot of Knowledge" or "Expert Knowledge" in logic and programming, we expect to see accuracy increase with knowledge once data collection is complete. However, even within the group of participants with "No knowledge", many are performing at or near ceiling. This can be seen in the plot below, where gray dots represent each participant's mean accuracy and reaction time.

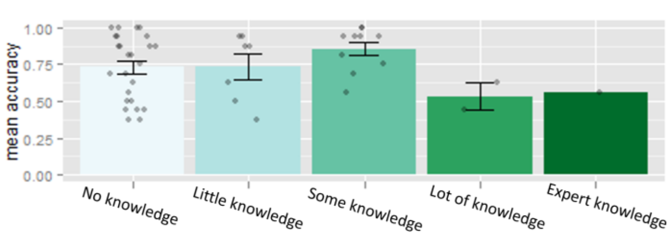

Fig. 3. Mean accuracy by logic level (and worker) with SE

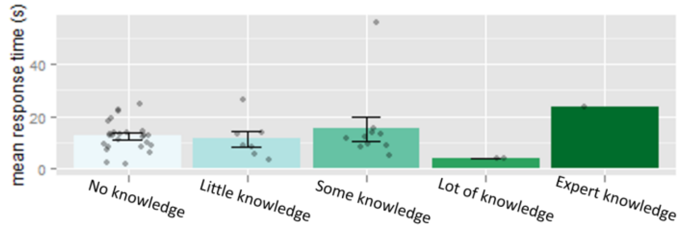

Fig. 4. Mean reaction time by logic level (and worker) with SE

Below, the contrast of interest between the uniqueness expressions is shown, where gray dots represent each participant's mean accuracy and response time for each type of uniqueness expression. A generalized linear mixed model with worker and rule form as random effects revealed no effect of uniqueness expression on accuracy ($\chi^2(2)$= 0.615, p=0.735) or response time ($\chi^2(2)$=0.555, p=0.758).[6][7]

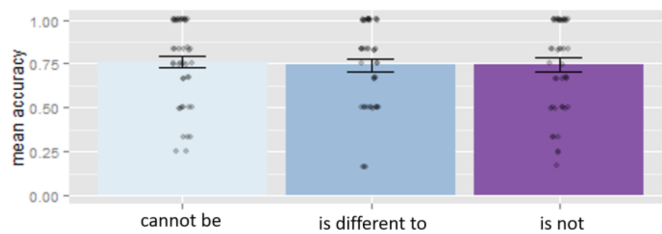

Fig. 5. Mean accuracy by uniqueness expression (and worker) with SE

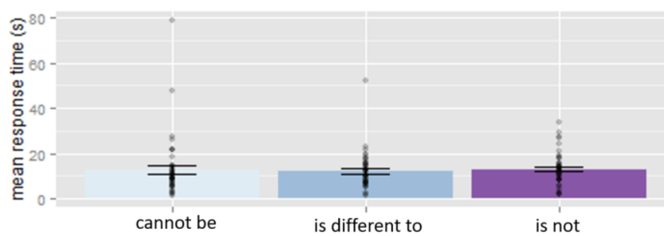

Fig. 6. Mean response time by uniqueness expression (and worker) with SE

D. Discussion of MTurk approach

The relative high accuracy attests that this paradigm works, allowing participants to demonstrate comprehension of CE. Notably, while some participants' accuracy is near chance (0.5), others are performing at or near ceiling, even within the group reporting "No Knowledge". This suggests that a background in logic or programming is not a prerequisite to success at this particular task, or perhaps to CE tasks in general. Furthermore, while no significant effect of uniqueness expression was found, an effect may have been masked within those performing at ceiling.

Response times point to two issues. First, in figure 2 several participants (those with low mean accuracy) have very short response times, suggesting that they did not read the CE rule and inspect the diagram before responding. In future analyses, these data points can be removed. Second, inspection of individual trials reveals a number of response times well above 30 seconds. These longer response times, combined with high overall accuracy, suggest that a time limit may lower performance and may be crucial for identifying performance differences within CE. As data collection progresses and various hypotheses are tested, we will be able to make evidence-based

recommendations for the design and use of CE and other Controlled Natural Languages.

IV. Analytic Pitfalls and Assumptions

A key aspect of supporting human users in problem solving is the avoidance of analytic pitfalls, and this was explored in the work on the Analysis Game [2]. As this work provided a model for the design and interpretation of the face-to-face style initiative, we provide a brief summary of the work undertaken.

The Analysis Game is a logic problem taught to analysts to explore how analytic reasoning can fail, being subject to "pitfalls" such as "layering" (where assumptions are used to come to conclusions, but the assumptions are then forgotten) and "mirroring" (where it is presumed that other people have the same concepts and world viewpoint as the analyst). This logic problem was modelled in CE, and collaborative reasoning between man and machine was employed to successfully solve the problem. It was discovered [2] that a suitable way to achieve collaboration between man and machine was to proceed by a process of "iterative formulation", whereby the user iteratively builds up a conceptual model and tests it out, thereby receiving immediately feedback on its usability and correctness. This is diagramed below:

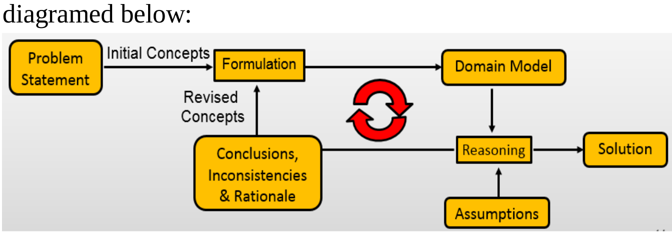

Fig. 7. Iterative formulation

In the CE formulation it was found that the making of assumptions was an important component, and that the use of assumptions could be exposed as part of the rationale thus helping to reduce analytic pitfalls. The design of the face-to-face evaluation followed much of the work on the Analysis Game: a logic problem was chosen, the Microsoft Word interface was used to support iterative formulation, and the use of assumptions by the users were an important issue.

V. The Face-Face Evaluation

A. Design principles

The following design principles [4] were developed:

- We aim to explore how CE could assist an analyst in the formulation and execution of a simple problem solving task, combining the requirement to formulate the problem in CE, together with the development of rules to infer the solution.
- The setup will permit iterative formulation; the system will run the reasoning and show the results and rationale.
- The user will only be asked to solve part of the problem, and some of the conceptual model and some facts will be predefined.
- The user will be introduced to the basic ideas of CE, including the predefined conceptual model and how facts may be expressed in sentences
- Tooling will be provided to assist the user, supported by the evaluator. The user will be expected to read and understand the CE without intervention, and will be expected to formulate CE concepts and facts, although the input of such information may require help from the evaluator.
- This evaluation is not yet intended to be a formal experiment with control groups to measure and compare performance; the purpose is to obtain informal feedback on how users use CE for problem solving and how a more formal test could be constructed in the future.

B. The "Three-ties" problem

We chose to use a simple logic problem as the basis of the study task, as this provides a simple entry into devising techniques that could be used for the study, whilst affording sufficient complexity to require the users to think about the problem and be creative about its formulation.

The logic problem chosen is the "Three Ties" problem created by Martin Gardner. It is relatively simple, but contains a "twist" that requires common sense knowledge to solve it; once formulated, the CE reasoning can easily come to a conclusion. It is a simpler example of a type of problem exemplified by the "Analysis Game". The text of the problem is as follows:

An Englishman (Mr Salmon), a Welshman (Mr Green), and a Scotsman (Mr Brown) met for lunch one day. One man was wearing a salmon tie, another was wearing a green tie and the third was wearing a brown tie. "Isn't it funny," said Mr Brown to the others, "that not one of us is wearing a tie which matches our name?" "That's true," agreed the man wearing the green tie. Can you now say what colour tie each man was wearing?

C. Training

An informal training session was held before any tasks were attempted, to show users the basic ideas of CE, and provide a "cheat" sheet of the predefined concepts. The user was asked to work with a system that can perform reasoning and deduction, but that "one vital piece of information has been omitted" and it was the user's task "to look at some of the sentences and express them to the system, so that the problem can be solved".

The training slides are provided in [4], and the "cheat sheet" is reproduced below to give an idea of the concepts involved:

Types	Known things	Descriptions of things
the man	MrGreen, MrBrown, MrSalmon	GreenTieWearer, BrownTieWearer, SalmonTieWearer
the tie	GreenTie, BrownTie, SalmonTie	MrGreensTie, MrBrownsTie, MrSalmonsTie
What you can say about things…		and the opposite…
the man X is wearing the tie Y.		the man X cannot be wearing the tie Y.
the thing X is the same as the thing Y.		the thing X cannot be the thing Y.

This table shows a key distinction that is not obvious from the normal reading of the logic problem, the distinction between specific, identifiable things (e.g. MrGreen, GreenTie) and descriptions of things that have not been identified (e.g. GreenTieWearer – the person wearing the green tie whoever that

may be). This distinction is key to the solution of these types of problems [2]

D. Evaluation Setup and Tasks

Five users undertook the evaluation, and one of the authors acted as evaluator and facilitator. Most users were familiar with programming, as well as having some passing knowledge of CE, although none had actually tried to use CE to formulate a problem. These five users undertook the series of tasks, but only four completed the series. All of the users gave permission for the evaluator to record their responses, and all expressed interest in issues that arose from the attempts to solve the problems.

The "Three Ties" logic problem was split into three tasks:

- **Warmup task** to familiarise the user with the setup and the writing of CE; this requested a formulation of one of the sentences "*"Isn't it funny," said Mr Brown to the others, "that not one of us is wearing a tie which matches our name?"*
- **Basic task** to consider the implications of one of the sentences *"That's true," agreed the man wearing the green tie.* and to formulate these implications as one or more CE sentences, leading to the solution of the problem
- **Advanced task** to explore the construction of new concepts and rules, by requesting the user to provide a more detailed explanation of the CE sentence(s) derived in the basic task. As this is a complex task, involving knowledge engineering, a process was devised to lead the user though several stages via a form: constructing a free English explanation; identifying nouns and verbs that might form new concepts underlying the explanation; generating a "because" statement linking user-devised CE-style premise sentences with the actual CE conclusion fact (from task 2); converting the "because" sentence into a rule and the "CE-style" premises into "conceptualise" statements.

For each task an "evaluator's form" was created, providing a set of questions to be asked of the user by the facilitator, together with space to record the user's answer. In this way there was a guided sequence of activities that led the user to a possible formulation of the problem. In the advanced task there is a significant component of "knowledge engineering" where users had to devise a logical representation of the knowledge, irrespective of whether CE is to be used, and it was thought necessary to provide the user with guidance of how this was to be performed. An example form is given below [4]:

1. **Write down in free form English what this phrase means for the individuals involved and the ties they can or cannot be wearing …**
2. **Write CE facts that capture this information[1]. The facts from the warmup exercise are already given:**

the man MrBrown cannot be wearing the tie BrownTie. the man MrGreen cannot be wearing the tie GreenTie. the man MrSalmon cannot be wearing the tie SalmonTie.

3. **Run the App by pressing the "Submit" button, and check the answer in the tables below**

CANNOT WEAR	*GreenTie*	*BrownTie*	*SalmonTie*
MrGreen	the man MrGreen cannot be wearing the real tie GreenTie.		the man MrGreen cannot be wearing the real tie SalmonTie.
MrBrown	the man MrBrown cannot be wearing the real tie GreenTie.	the man MrBrown cannot be wearing the real tie BrownTie.	
MrSalmon		the man MrSalmon cannot be wearing the real tie BrownTie.	the man MrSalmon cannot be wearing the real tie SalmonTie.

the man MrBrown is wearing the real tie SalmonTie. the man MrGreen is wearing the real tie BrownTie. the man MrSalmon is wearing the real tie GreenTie.

4. **What did you find easy and what did you find difficult? …**
5. **Can you provide an explanation for the thinking you did in the box about "what this phrase means for the individuals involved"? …**

These forms perform two functions. Firstly they act as a repository for the answers by the user (together with any prompting or questioning by the evaluator), which provide the raw material for the analysis in the "Discoveries" section. Secondly, they are "active" in that the CE formulations provided by the user can be automatically compiled and executed by the CE reasoning system, and the results can be viewed in the query tables also embedded in the forms. Thus it is not necessary to have a separate interface for the CE reasoning tool and the record of the questions and the answers.

E. Discoveries from the Face-to-Face records

We have undertaken an informal analysis [4] of the user's responses recorded in the evaluation forms; these include informal thoughts, explanations, etc., as well as the formal construction of CE sentences and rules. In this section we present this analysis under several key topics, and include some short extracts from the forms, flagged between double quotes; the text in the forms was actually typed in by the evaluator, but was simultaneously checked and approved by the user. No subsequent attempt has been made to re-edit this, although some additions [in square brackets] have been made by the author to fill in missing words.

1) Turning English into Controlled English

The tasks encouraged the users to write their knowledge into free, unconstrained English as a step towards formulating CE sentences. All users perceived the essential nature of the information in the first two tasks, involving people and the ties they were or were not wearing, but there was variability in the linguistic expressions used in free English, for example "*Mr*

[1] e.g. the man MrBrown cannot be the man GreenTieWearer.

Brown cannot be wearing/is not/is'nt wearing the Green tie" or "*Mr Brown is not the green tie wearer*".

There were two main ways that the formulation occurred, one based upon the relationship between a person and their tie, such as "*the person MrBrown cannot be wearing the tie GreenTie*", and the other based on the comparing of descriptions with specific individual entities, such as "*the person MrBrown cannot be the person GreenTieWearer*". There were also some variations between the use of "man" and "person". Thus there is variability in the formulations, but in the first two tasks this was limited and constrained by the set of expressions predefined in the CE model. Far more creativity was evident in other factors that were captured in the records of tasks, as described below.

2) Further inferencing and creativity

When users were asked to formulate specific sentences into CE, many of them extended their thinking beyond the information contained in the sentences, and started to generate further inferences from the information. For example, some users inferred the next step: "*Mr Brown cannot be wearing the green tie*"; some suggested that a man could be wearing two of the ties: "*or mr salmon could be wearing the green or the brown [tie]*". One user went further and concluded that "*the person MrSalmon is wearing the tie GreenTie*". Creativity extended beyond this inferencing: one user raised (and later rejected) the possibility that the adjective "Green" might be a brand name or even a type of tie. All of these inferences are correct, but are not directly stated in the sentence being formulated, thus formulation is not separated from problem solving.

3) Assumption making

The richest creativity was shown in the making of assumptions to focus in on the most probable account of the situation described in the logic problem. Such assumptions were varied, creative, and spontaneous. A key assumption that was made explicitly by most users is that Mr Brown is not talking to himself (and is thus indicative that he is not the same person as the green tie wearer). There were many different ways that users devised that assumption, for example: "*assuming MrBrown isn't talking to himself, it can't have been MrGreen replying as the responder is wearing a green tie*", "*Assumption is that there are people talking to each other and that there are two people in that interaction*", "*assumption is that the answer is by someone other than mrbrown.*", "*but he could be schizophrenic. This is how an author might express split personality*". These recorded statements (in the free English sections of the forms) suggest that assumption making is part of constructing an explanation. Some assumptions were more creative: "*assume greentie wearer is not colour bind and has visibility of all ties and wearers*".

Key assumptions that were not made by most users is that a tie cannot be worn by more than one person, and that one person cannot wear more than one tie. The problem cannot be solved unless these assumptions are made, but only one person made such reasoning explicit, albeit in an obscure way: "*assumes that the properties cannot be shared by individual people. property [here] is the colour of the tie.*"

Several "triggers" seemed to encourage the making of assumptions: when formulating the CE sentences and rules assumptions were spontaneously created; when explaining the reasoning, assumptions were also spontaneously created; when questioned on alternative possibilities some users were able to create new possibilities, and then erected assumptions to discount them.

4) Exposing the problem-solving process

Users were encouraged to work through the problem in stages, generating free English statements, then CE formulations and finally providing an informal explanation of the reasoning they undertook. This seems to have had the effect that their reasoning process was captured and made explicit. The processes involved in formulating sentences and in making assumption have already been described, but further stages of the problem solving were also captured.

At the end of the first task, users were asked to describe missing information that was needed to solve the task. All noticed that an extra constraint was necessary (and some made reference to the "cannot wear" table, which was only populated one person per tie), but the way they expressed this was variable (and creative): *"[need] to eliminate more options*", "*one more constraint (e.g. if a tie is known for one of the men then the rest will fall into place*". Two users intuited that the answer might lie in the next sentence: "*is there any information from the fact that the man from the green tie agreed with Mr Browns statement?*".

Explanations for the state of the reasoning were spontaneously (and occasionally after prompting) generated. Some of these explanations were contained as part of making an assumption, as described above. Explanations were also made in their own right; some were basic: "*mr brown is not wearing the green tie as somebody else is*" (Note that this also embodies the assumption that only one person can wear the green tie, though this is not made explicit). Some were deeper: "*it's a conversation between MrBrown and someone else. The other person is wearing a green tie. therefore mr brown isn't.*". These were all provided in the first two tasks. The third task encouraged the construction of deeper explanations in order that new concepts could be created to formalise the reasoning at a greater level of detail; examples are given below.

For several users, the working of the reasoning towards a solution was evident: "*it doesn't say that mrbrown isn't the same person as the green tie wearer. could be that mrbrown and the man wearing the green tie are the same. most likely situation is that mrbrown is not the green tie wearer, [ah!] therefore the person MrBrown cannot be wearing the tie greentie.*".

Several users noted that knowledge about the genre of the problem itself, as context, affected their reasoning and formulation: *"[the red sentence was a] clue to [the problem]. knowing it's a puzzle influences the way I thought about it*".

5) Reaction to machine problem solving

Several users reported surprise when the solution was auto-generated after the provision of the extra information in the second task: "*Initially surprised that the answer was arrived at so easily [by the system]. Then reviewed the matrix and worked it out how the answer had been obtained*", "*this was impressive that it did the unravelling*".

6) Different styles of reasoning

Even with the small number of usersdifferent styles of reasoning were involved across different users. One notable

difference was exhibited by one user who quickly jumped to the inference that "*the person MrSalmon is wearing the tie GreenTie*" (noted above), and thereby almost came to the solution without the use of the CE system. However when it came to the third task, it appeared difficult for that user to let go of that specific inference, and it proved difficult to construct the abstract generalization needed to provide a deeper explanation.

7) Creating new concepts and new rules

The third task requests the user to construct a deeper explanation of the reasoning for the CE sentence formulated in the second task (e.g. "*the person MrBrown cannot be the person GreenTieWearer*"). In this section we call this sentence the <CONCLUSION> and we are seeking an explanation in the form *"<CONCLUSION> because <PREMISE>"*, where the user is to provide a CE sentence to be the <PREMISE>, for example "*the person MrBrown cannot be the person GreenTieWearer because the man GreenTieWearer agrees with the man MrBrown*". This explanation will use concepts that are not yet in the conceptual model, and the user is assisted by the experimental knowledge engineering approach noted above.

Most of the users were able to identify the concepts underlying deeper explanation. They were encouraged to identify nouns and verbs (leading to concepts) and whilst they did identify simple concepts such as "tie", "colours", "two people", in practice the informal conceptualisation came with more detail: "*'not one of us' means is not a specific group of people we have been told about*", "*conversation between two people; normally a conversation is between two people*", "*quotes separating the sentences from the rest of the narrative give a clue that it is a conversation; attaching bits to the person saying it*". The free English explanation could also be easily expanded: "*It's this (agreed) that indicates a different person. Eg "continued" would indicate that it was the same person. But not necessarily, but more likely*".

The next step involved the creating of an explanation in the form *"<CONCLUSION> because <PREMISE>"*. The user first wrote <PREMISE> in free English and then in a new CE-style sentence <u>devised by the user</u>. The "CE-style" sentence is not yet CE since the relevant concepts are not formally conceptualized, but they allow the user to write a formal sentence that could be turned into a CE sentence (after the concepts are pulled out of the sentence and defined). Examples of free and CE versions of <PREMISE> <u>written by the user</u> (separated by /) are: "*the man wearing the GreenTie agrees with MrBrown*" / "*the man GreenTieWearer agrees with the man MrBrown*"; "*there was a conversation between two people and one of them is wearing the green tie and the other person in the conversation is Mr Brown*" / "*the thing X is conversing with the thing MrBrown".*

The "because" sentence can be turned into a new conceptualisation for the CE-style PREMISE and an "if-then" CE rule, after the specific instances of things are turned into variables. This rule is a re-useable component that is more generic than the original "because" sentence, and can play a part in the construction of inferences and rationale. For example, the "because" sentence noted above can generate the following concept and rule:

conceptualise the man X ~ agrees with ~ the man X1.

if (the man A agrees with the man B)
then (the man B cannot be the man A).

Four of the users were evaluated on this third task, and all managed to generate "because" sentences, CE-style sentences, and CE rules, facilitated by the evaluator. One of the rules had a problem, caused by an obscure detail about how the CE interpreter was implemented, which requires a code modification to warn the user of this problem. The rules generated by the other three users were successfully employed into the CE model. For these users, it was possible to demonstrate rationale graphs that showed their rules, and this provided confirmation that the rules were operating correctly.

8) Ease and difficulties

In all three tasks, the users were asked the general question of what things were easy and what things were difficult. No specific question about CE was asked, and all the comments in this section were spontaneous. In the first two tasks, no users reported any significant difficulties in the formulation of the free English sentences into CE, although some users checked against the "cheat sheet" to see the possible expressions. Some comments were: "*CE seemed obvious as a way to express it*" (from a user who had previous exposure to CE); "*easy to do short Controlled English sentences*"; "*writing CE relatively easy, refer to cheat sheet, easier that the previous statements were present. problem would be difficult if it wasn't there, and I wouldn't know what words to use*".

In the first two tasks, some general difficulties in formulation (unrelated to use of CE) were reported: *"in getting the 'agrees' sentence into a logical expression." ; "difficult to do the flip to make it about the person not the tie"*. (This was by the person who was "fixed" on jumping ahead to formulate that Mr Salmon was wearing the green tie).

For the first two tasks, the use of CE is partly about selecting suitable phrases from the subset defined in the cheat sheet, whereas in the third task more creative effort was needed to construct suitable sentences and concepts. However, in the third task the use of CE was still found to be relatively easy: "*actually CE is quite easy. takes a bit of getting used to.*", "*easy to write the sentence once a conversation was [conceptualised as] a relationship*. " (since the evaluator was typing the CE, these comments refer to expression of concepts rather than typing).

In the third task there were issues in defining the right logical concepts (relating to conceptualisation rather than CE):"*tricky to get down to the basic concepts, there was assumed knowledge that I didn't realise I was using … trying to jump from puzzle to CE without getting the intermediate steps could end up with lots of duff rules.*"; "*there was a question as to whether [the conversation] was a relationship or a thing.*"

9) A mistake by the evaluator

As noted above, one user focused on the inference that MrSalmon was wearing the green tie, rather than the fact that the GreenTieWearer was not MrBrown, which had been the focus of the other users. The user did not make explicit the first inference (the two people are not the same) and built an explanation "MrSalmon was wearing the green tie because there was a conversation between two people". The evaluator pointed out the strangeness of this explanation and prompted the user to

move to a "more reasonable" explanation. It would have been more interesting to let the user run with this original explanation, to the point at which a general rule was to be generated, to see if the strangeness of the explanation was noticed and revised.

F. Future evaluations

The basic evaluation structure seems useful; it follows the design principles and reveals insights into cognitive reasoning and effects of CE. However some improvements are:

- The first and second tasks seem too simple and a more challenging problem should be constructed. Since some users reasoned on the basis that it was a logic problem, the tasks should seem less like solving a "logic puzzle".
- In this problem the "common sense" assumptions are correct and it is unlikely that a user will follow an incorrect assumption. Therefore a "garden path" problem should be devised where the common sense assumptions are incorrect, leading the user to make incorrect assumptions, which then have to be revised.
- The evaluator should provide less guidance to users, and the roles of facilitator and evaluator should be separated
- The evaluation analysis focuses on "horizontal" comparisons between users solving the same steps in a task, whereas it would be interesting to understand specific styles of reasoning by a "vertical" analysis that traces down an individual's progress towards the conclusion.

G. Discussion of Face-to-Face evaluation

The observations described above lead to thoughts about the users' cognitive behaviour and the effects of using CE. It is not claimed that these are formally validated, but they are offered as interpretations that can serve as the basis for future experiments.

Even in the relatively simple first and second tasks, where it would seem that there were limited options for formulation, the users were highly creative and utilized "common sense" and this was strongest in the third task. Creativity was exhibited in selecting different CE formulations, in making assumptions and in the construction of explanations and new concepts.

Users naturally seemed to infer more information than was directly contained in the sentence to be formulated, and these inferences seemed to drive towards solving the problem. It would be interesting to see if such secondary inferencing occurs when solving a problem is not the goal.

Assumptions are made fluently by users, either spontaneously, or after prompting. However more fundamental assumptions were not generally stated, such as that only one tie was worn by one person; noted by one user. There was some evidence that this assumption was actually being used though not explicitly, so it is possible that "obvious" assumptions do not float to the surface of consciousness. However expression and communication of assumptions is key to avoiding analytic pitfalls, so this is a key area to research further.

The evaluation recorded some detail about the user's problem solving and conceptualisation strategies, together with issues that they faced, although it should be questioned whether what was recorded was actually the reasoning process or the user's perception of the reasoning process after being turned into language for the purpose of communication with the evaluator. However users did seem to perform reasoning and conceptualisation in different styles. For example, one user was very focused on following a specific line of reasoning, and seemed reluctant to abstract away from this specific line; it seems unlikely that this was just an artifact of the evaluation.

The collaborative Word document interface used in the evaluation allowed the users to input CE facts and to see how this led to a solution to the problem. Such an interface is atypical of a computer system and a Word document, and some users were surprised and impressed by the behaviour of the system.

Generally, users were able to formulate the problem in CE and construct new CE concepts to model the world and provide explanation. Use of CE did not lead to specific difficulties, over and above the problems associated with conceptualisation of the world in a formal way. There was some informal evidence that the use of CE (and the use of the "cheat sheet") did provide tools for some users in the construction of formulations.

VI. CONCLUSIONS

Both approaches provided some evidence that users were able to understand CE sufficiently to answer logical questions and to use it for formulating sentences and creating conceptual models that could be used to solve simple problems. Although the two approaches were different, they were complementary and both tested the same language, and results in one approach could be relevant to the other. Both approaches are still at the early stages, but the results suggest that both provide valid experimental frameworks with the potential to learn more about CE and the issues involved in its use by humans. It is intended that both of these frameworks serve as the basis for further work.

ACKNOWLEDGMENT

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Mott, D. (2010) Summary of CE, https://www.usukita.org/papers/5658/details.html

[2] Mott, D., Shemanski, D.R., Giammanco, C. ,Braines, D., Collaborative human-machine analysis using a Controlled Natural Language, SPIE Next-Generation Analyst III, April 2015, https://www.usukitacs.com/node/2852.

[3] Tobias Kuhn. An Evaluation Framework for Controlled Natural Languages. In *Proceedings of the Workshop on Controlled Natural Language* (CNL 2009). Springer, 2010

[4] Mott, D., Evaluations of Controlled English, May 2015, https://www.usukitacs.com/node/2954.

[5] Kuhn 2010 Controlled English for Knowledge Representation, Doctoral thesis, University of Zurich

[6] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[7] Bates D, Maechler M, Bolker B and Walker S (2014). *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-7, http://CRAN.R-project.org/package=lme4

1 (PDF) DEFENSE TECHNICAL
INFORMATION CTR
DTIC OCA

2 (PDF) DIRECTOR
US ARMY RESEARCH LAB
RDRL CIO LL
IMAL HRA MAIL & RECORDS
MGMT

1 (PDF) GOVT PRINTG OFC
A MALHOTRA

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM D
T DAVIS
BLDG 5400 RM C242
REDSTONE ARSENAL AL
35898-7290

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRS EA DR V J RICE
BLDG 4011 RM 217
1750 GREELEY RD
FORT SAM HOUSTON TX
78234-5002

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM DG K GUNN
BLDG 333
PICATINNY ARSENAL NJ
07806-5000

1 (PDF) ARMY RSCH LAB – HRED
ARMC FIELD ELEMENT
RDRL HRM CH C BURNS
THIRD AVE
BLDG 1467B RM 336
FORT KNOX KY 40121

1 (PDF) ARMY RSCH LAB – HRED
AWC FIELD ELEMENT
RDRL HRM DJ D DURBIN
BLDG 4506 (DCD) RM 107
FORT RUCKER AL 36362-5000

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM CK J REINHART
10125 KINGMAN RD BLDG 317
FORT BELVOIR VA 22060-5828

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM AY M BARNES
2520 HEALY AVE
STE 1172 BLDG 51005
FORT HUACHUCA AZ
85613-7069

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM AP D UNGVARSKY
POPE HALL BLDG 470
BCBL 806 HARRISON DR
FORT LEAVENWORTH KS
66027-2302

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM AR J CHEN
12423 RESEARCH PKWY
ORLANDO FL 32826-3276

1 (PDF) ARMY RSCH LAB – HRED
HUMAN SYSTEMS
INTEGRATION ENGR
TACOM FIELD ELEMENT
RDRL HRM CU P MUNYA
6501 E 11 MILE RD
MS 284 BLDG 200A
WARREN MI 48397-5000

1 (PDF) ARMY RSCH LAB – HRED
FIRES CTR OF EXCELLENCE
FIELD ELEMENT
RDRL HRM AF C HERNANDEZ
3040 NW AUSTIN RD RM 221
FORT SILL OK 73503-9043

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM AV W CULBERTSON
91012 STATION AVE
FORT HOOD TX 76544-5073

1 (PDF) ARMY RSCH LAB – HRED
RDRL HRM DE A MARES
1733 PLEASONTON RD BOX 3
FORT BLISS TX 79916-6816

8 ARMY RSCH LAB – HRED
(PDF) SIMULATION & TRAINING
TECHNOLOGY CENTER
RDRL HRT COL G LAASE
RDRL HRT I MARTINEZ
RDRL HRT T R SOTTILARE
RDRL HRT B N FINKELSTEIN
RDRL HRT G A RODRIGUEZ
RDRL HRT I J HART
RDRL HRT M C METEVIER
RDRL HRT S B PETTIT
12423 RESEARCH PARKWAY
ORLANDO FL 32826

1 ARMY RSCH LAB – HRED
(PDF) HQ USASOC
RDRL HRM CN R SPENCER
BLDG E2929 DESERT STORM DR
FORT BRAGG NC 28310

1 ARMY G1
(PDF) DAPE MR B KNAPP
300 ARMY PENTAGON
RM 2C489
WASHINGTON DC 20310-0300

ABERDEEN PROVING GROUND

12 DIR USARL
(PDF) RDRL HR
L ALLENDER
P FRANASZCZUK
K MCDOWELL
RDRL HRM
P SAVAGE-KNEPSHIELD
RDRL HRM AL
C PAULILLO
RDRL HRM B
J GRYNOVICKI
RDRL HRM C
L GARRETT
RDRL HRS
J LOCKETT
RDRL HRS B
M LAFIANDRA
RDRL HRS D
A SCHARINE
RDRL HRS E
D HEADLEY
E ZAROUKIAN